Databases and Data Mining

Databases Assignment 3

Due: Wednesday 28-11-2012
Grading: This assignment will be graded from 0 to 10.
Notes:

- Groups of 1-2 students are allowed.
- Use the Weka system to mine the association rules as well as for preparing the data and presenting the results.
- Write down your technical report for this assignment in a *.pdf* file with the following name **"<your student number><your name>_3.pdf"**, e.g., **"012345janjansen_3.pdf",** or **"012345janjansen_678910_ansjansen_3.pdf"** if you are working in a team of 2.
- Send this .pdf file as an attachment of an e-mail with subject **"DBDM_3"** to *erwin@liacs.nl* .
- Do not use more than 8 A4 (font size 10 pt) for your report.
- Grading will be based on
  - the quality of your data mining strategy and results
  - the argumentation, validity, and clarity of your report.

Introduction

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. For example, the rule *{onions, potatoes}* → *{beef}* found in the sales data of a supermarket would indicate that if a customer buys *onions* and *potatoes* together, he/she is likely to also buy *beef*. In addition to the above example from market basket analysis, association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.
In this assignment you have to use the association rule mining module of the Weka system to mine association rules from the *Adult Data Set*.

Dataset

The *Adult Data Set* [1] was originally extracted from the census bureau database found at http://www.census.gov/ . The data set contains 14 attributes such as age, work class, race, sex, etc. along with an indication of whether or not that person makes over 50K a year. The main prediction task was to determine whether a person makes over 50K a year. In this assignment you are ***also*** asked to find other interesting rules. You can download the *Adult Data Set* and find more detail information about the dataset using one of the following links:

http://archive.ics.uci.edu/ml/datasets/Adult (obtain the data by following the *Data Folder* link on this page)
http://www.sgi.com/tech/mlc/db/ (only use this link, if the previous link did not work)

WEKA

Weka is a collection of machine learning algorithms for data mining tasks that contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. All the information (i.e., software, documents and book) you need about *Weka* can be found using the following link and/or reference [2]: http://www.cs.waikato.ac.nz/~ml/weka/index.html

Guidelines

- Due to the representation of frequent itemsets in Weka, this system may run out of memory when mining datasets with as few as a dozen attributes. Run several experiments with your data and the system varying the parameters until you obtain a collection of association rules that represent your data well.
- Use the Weka system to mine the association rules as well as for preparing the data and presenting the results. Code by yourself any functionality that you need for manipulating the data, and any necessary functionality that is not offered in the Weka system.
- You can restrict your experiments to a subset of the dataset if Weka cannot handle the whole dataset. But you should take care to reduce the probability of missing any of the representative association rules while mining the data.
- After you have cleaned and selected a subset of your data (if necessary), mine association rules using different parameter (confidence, support, etc.) settings. Analyze the resulting rules and repeat the experiment with another ”view” of the data given by generalizing/specializing your data according to the concept hierarchies and/or by selecting different portions of the data.
- Use explicit pattern evaluation measures to evaluate the quality of your discovered patterns. Use for example the Kulzcynski measure in combination with the imbalance ratio, etc. (see slides of the 6-11 lecture). Explain your choices!
- Assume that you as the user/miner want to obtain association rules for decision support, for understanding the data better, and/or for increasing your company’s profit. Mine rules until you obtain a collection of rules that satisfy these objectives.

Report

Your report should start with a *title*, the *names and student numbers of the team*, an *abstract and introduction*, and at least contain the following sections with the corresponding discussions:

Statistical report

- Report the mean, median, minimum, maximum and standard deviation for each of the numerical variables.

Code Description

- Describe the code that you used/wrote. Remember to acknowledge any sources of information/code you used.

Experiments:

- Describe in each case the objectives of your analysis: Is it to understand the data better? If so, what is it about the data you want to understand? Or is it for decision support? If so, what decisions you need to make based on the data?
- For each experiment you should describe:
    - Instances: What data did you use for the experiments?
    - Any pre-processing done to improve the quality of your results.
    - Your system parameters.
    - Any post-processing done to improve the quality of your results.
    - Analysis of results of the experiment and their significance.

Summary of Results

- What was the best collection of association rules that you obtained? Describe. Discuss the strengths and the weaknesses of your mining and evaluation methods. What are your conclusions?

References

[1] Ron Kohavi, *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

[2] Daniel T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Chapter 3, John Wiley, Chichester, 2004. (See also: http://dataminingconsultant.com/DKD.htm .)

[3] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Chapter 9, Morgan Kaufman, 2005.